Show-through Cancellation and Image Enhancement by Multiresolution Contrast Processing

Alicia Fornés, Xavier Otazu and Josep Lladós
Computer Vision Center, Dept. of Computer Science
Universitat Autònoma de Barcelona, Edifici O
08193, Bellaterra, Spain
{afornes,xotazu,josep}@cvc.uab.es

***Abstract*—Historical documents suffer from different types of degradation and noise such as background variation, uneven illumination or dark spots. In case of double-sided documents, another common problem is that the back side of the document usually interferes with the front side because of the transparency of the document or ink bleeding. This effect is called the show-through phenomenon. Many methods are developed to solve these problems, and in the case of show-through, by scanning and matching both the front and back sides of the document. In contrast, our approach is designed to use only one side of the scanned document. We hypothesize that show-trough are low contrast components, while foreground components are high contrast ones. A Multiresolution Contrast (MC) decomposition is presented in order to estimate the contrast of features at different spatial scales. We cancel the show-through phenomenon by thresholding these low contrast components. This decomposition is also able to enhance the image removing shadowed areas by weighting spatial scales. Results show that the enhanced images improve the readability of the documents, allowing scholars both to recover unreadable words and to solve ambiguities.**

I. Introduction

The recognition of historical documents implies dealing with the degradation of paper due to paper aging. Typical types of degradation and noise include background variation, uneven illumination or dark spots. In case of double-sided documents, a common problem is that the back side of the document usually interferes with the front side because of the transparency of the document or ink bleeding. This effect is called Back-to-front or Show-through, which is more or less severe depending on the porosity and translucidity of the paper, the kind of the ink (percentage of iron) and the paper age.

Some existing state of the art methods binarize the documents and remove the show-through components [1], [2], [3]. However, our final goal is not binarizing these documents. Firstly, because many current handwriting recognition systems can work directly with grey-scale documents [4]. Secondly, because our aim is to enhance the document images in order to improve the readability of these documents, allowing scholars both to recover unreadable words and to solve ambiguities.

Many methods remove the show-through components by scanning and matching both the front and back sides of the document [5], [6], [7]. As expected, the performance of these methods depends on the accuracy of the matching of both sides. However, a perfect matching is not always possible: First, because of different skews and resolutions during scanning, second, warping surfaces because of thick bound books, and third, the verso can not be available (specially in historical documents, where the original document paper can be lost).

For these reasons, some approaches only use one side of the scanned document. For example, Wang et al. [8] remove the show-through components assuming that the handwriting is highly slanted. For this reason, the handwriting from the front and reverse sides can be detected using directional wavelet transforms at 45 and 135 degrees respectively. As expected, the authors state that this method can not deal with strokes of different orientations. Nishida and Suzuki [9] propose the restoration of color documents by estimating the background colors and correcting them through multi-scale analysis. However, this method seems to have problems when there are large show-through components. Lins, Silva et al. [10], [11] remove the framing borders, divide the image into blocks and classify the bleeding noise into three categories: weak, medium and strong. According to this classification, different threshold values are applied to each block of the document image. Then, they check the presence of blur in the image, and finally the show-through pixels are filled using linear interpolation. As the authors suggest, the performance of the method depends on the accuracy of the noise classifier to adjust the parameters.

In this paper we propose a show-through cancellation and document enhancement method that also uses one side. We hypothesize that show-trough are low contrast components, while foreground components are high contrast ones. A Multiresolution Contrast (MC) decomposition is presented in order to estimate the contrast of features at different spatial scales. Thus, we cancel the show-through phenomenon by thresholding these low contrast components. This decomposition is also able to remove shadowed areas by weighting spatial scales. Our main contribution is a method that only requires one side of the page, does not need to remove the framing borders, and since it only needs one parameter, it can be easily applied to any kind of document.

The experimental framework corresponds to old marriage records in microfilm format, which present a severe show-through effect. The experiments show that the enhanced images improve the readability of the documents, allowing scholars both to recover unreadable words and to solve ambiguities.

The rest of the paper is organized as follows. Section 2 is devoted to describe the Multiresolution Contrast method, whereas the experimental results are shown in Section 3. Section 4 concludes the paper and proposes future work.

II. Multiresolution Contrast Method

Our main hypothesis is that show-through are low contrast components, while foreground components are high contrast ones. Furthermore, we hypothesize that background variation are spatially wide components, hence they can be considered low spatial frequency features. Thus, our method has three main steps. First, we decompose the image into a Multiresolution Contrast representation [12], which allows us to obtain the contrast of components at different spatial scales. Second, we enhance the image by reducing the contrast of low spatial frequency components. Finally, we perform show-through cancellation by thresholding low contrast components. Following, we mathematically formalize these concepts.

A. Multiresolution Contrast Decomposition

The Multiresolution Contrast (MC) decomposition we present here is based in the combination of these two concepts: contrast and multiresolution decomposition.

Given image I, Michelson contrast is a measure of the contrast between two pixels. It is defined as

$$M_c(i,j) = \frac{I(i,j) - I(i',j')}{I(i,j) + I(j',j')} \ , \tag{1}$$

where (i,j) and (i',j') are the $(row, column)$ locations of the two pixels. Since it is a contrast measure, it is invariant to global ilumination changes (e.g. intensity of the scanned page, neighbour pixels on a shadowed area, etc).

A multiresolution decomposition of an image I can be defined as

$$I(i,j) = \sum_{s=1}^{n} d_s(i,j) + r_n(i,j) \tag{2}$$

where n is the number of spatial scales and d_s are the multiresolution coefficients at scale s, i.e. the details at scale s. A usual iterative procedure to obtain coefficients d_s is

$$d_{s+1}(i,j) = r_s(i,j) - r_{s+1}(i,j) \ , \tag{3}$$

with $r_{s+1}(i,j) = r_s(i,j) \otimes h_s$, $\otimes$ denotes convolution and h_s is a smoothing kernel. The term r_s is a smoothed version at scale s of original image $I(i,j)$. In the particular case where $h_{s+1} = \uparrow h_s$ ($\uparrow$ denotes diadic upsampling by inserting zeros), this decomposition scheme is equivalent to the *à trous* algorithm. It is the scheme we use in this work. We take

$$h_s = \frac{1}{256} \begin{pmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 8 & 24 & 8 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 8 & 24 & 8 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{pmatrix} \tag{4}$$

which leads to a biorthogonal wavelet decomposition.

In eq. (3), multiresolution coefficients are defined as an intensity *difference*. In contrast, we define Multiresolution Contrast coefficients as an intensity *contrast*.

We define Multiresolution Contrast coefficients by

$$\omega_s(i,j) = \frac{r_s(i,j) - r_{s+1}(i,j)}{r_s(i,j) + r_{s+1}(i,j)} \ . \tag{5}$$

This coefficients are a contrast measure of details at scale s, which implies they are invariant to global and local ilumination changes, where local means inside an area smaller than filter h_s.

Given these coefficients, the Multiresolution Contrast decomposition of image I (and its recovery from coefficients $\omega_s(i,j)$) is defined as

$$I(i,j) = r_n \prod_{s=1}^{n} \frac{1 + \omega_s(i,j)}{1 - \omega_s(i,j)} \tag{6}$$

B. Image processing

We perform both image enhancement and show-through cancellation by modifying the $\omega_s(i,j)$ coefficients of MC decomposition of the image.

We define the final processed image by

$$\hat{I}(i.j) = r_n \prod_{s=1}^{n} \frac{1 + \hat{\omega}_s(i,j)}{1 - \hat{\omega}_s(i,j)} \ , \tag{7}$$

where $\hat{\omega}_s(i,j)$ are the modified multiresolution coefficients. These coefficients are obtained by applying the following two steps:

1) Image enhancement: In the case of scanned documents, one of the main problems is background variation and uneven illumination. Since these defects cover relatively wide areas, they are present on the highest scales of the MC decomposition. Futhermore, MC allows to compensate for these problems while, at the same time, maintain the local contrast of spatially smaller features, e.g. text. In order to reduce the contrast of these wide defects we apply a weighting α_s function on the MC ω_s coefficients

$$\omega_s^{\alpha} = \alpha_s \, \omega_s \tag{8}$$

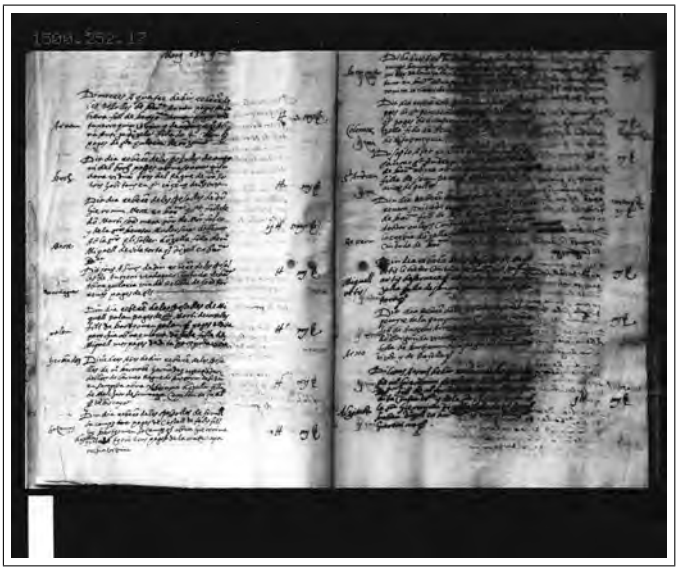
with $\alpha_s = e^{-\frac{s^2}{2\sigma^2}}$. Empirically, we obtain good results with $\sigma = 3$.

2) Show-through cancellation: We hypothesize that show-through are low contrast features. Hence, to remove them we perform a thresholding on the previously enhanced ω_s^{α}

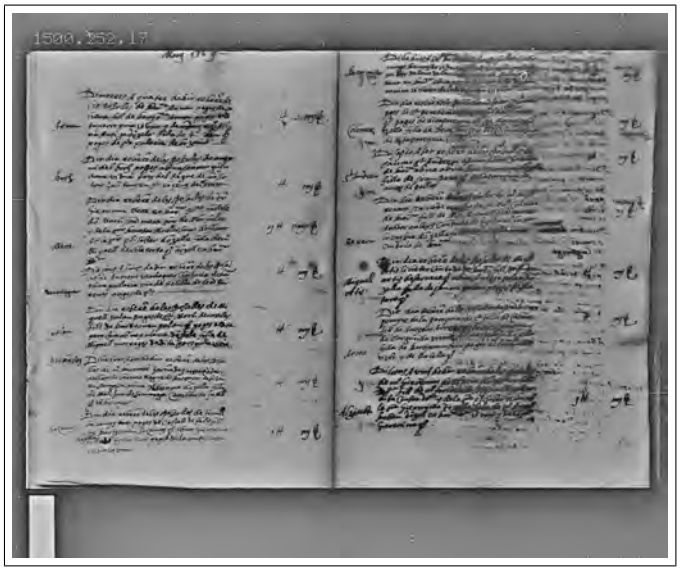
$$\hat{\omega}_s = thr(\omega_s^{\alpha}; \beta) \ , \tag{9}$$

being β the thresholding value.

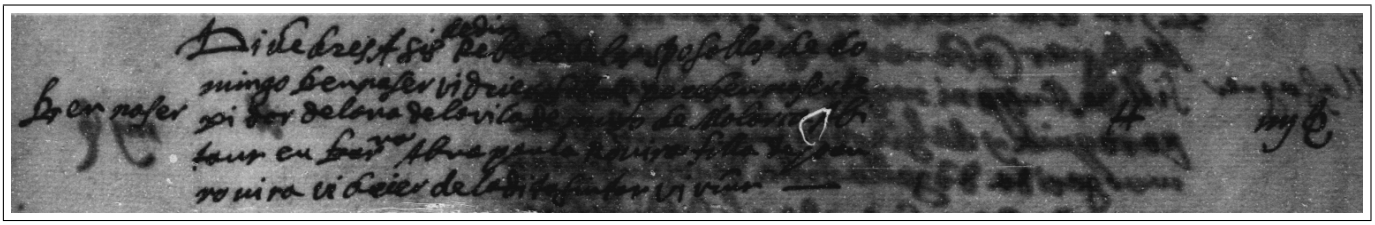
This final expression corresponds to the processed wavelet coeficients we use in eq. (7) in order to recover the final processed image.



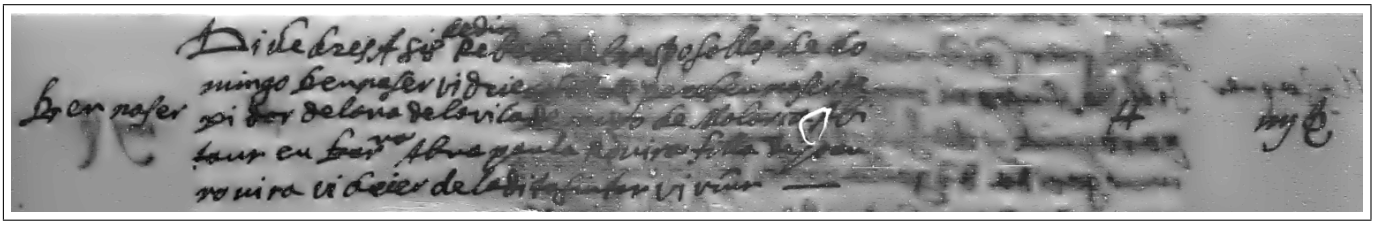
(a) Original Image



(b) Enhanced Image



(c) Section of the Original Image (corresponding to the top-right corner of 1a)



(d) Section of the Enhanced Image (corresponding to the top-right corner of 1b)

Fig. 1: Example of our document enhancement method applied to Microfilms. The show-through phenomenon is very severe at the right side page, where show-through components are very dark. Notice that it is not necessary to remove the framing borders.

III. Results

A. Dataset

The experiments have been mainly performed on the *Llibre d'Esposalles* [4]: a set of 244 books of marriage licenses records conserved at the Archives of the Barcelona Cathedral. They contain information of approximately 550.000 unions celebrated in over 250 parishes between 1451 and 1905. The registers contain information about the groom, bride, their parents, occupation, place of origin and the fee that was paid to the church. Information extracted from these manuscripts allows scholars not only to reconstruct genealogical trees, but also to study the demographical changes over five centuries.

From the 244 books, 21 books (around 150 pages per book) are only available in microfilm format. For our experiments, we have selected a subset of 22 microfilm pages, which were written in 1569. The microfilms are of a very bad quality (see Figure 1b), and since the original documents are lost, they can not be scanned again. For scholars, recovering as much information as possible is of key importance, because unreadable marriage registers will imply gaps in the genealogical trees.

In order to compare with other exsiting state of the art methods, we also show some qualitative results using document images extracted from [9], [10].

B. Metrics

Since we do not binarize the images, metrics proposed for assessing binarization algorithms [13], [3] can not be applied here. For this reason, in addition to the qualitative results, we propose quantitative readability measures, similar to the ones proposed in [8].

The procedure was:

1) The 22 original images were given to the scholar (in this case, an expert on historical documents), who extracted all the *important words* he could.
2) The 22 enhanced images were given to the scholar. He now reports the number of *important words* that were not readable at step 1 and he can read now (*new words*). Also, he reports the number of *important words* that were misreaded at step 1 and that he corrects at step 2 (*modified words*).

We define *Important words* as the ones that contain information (names, surnames, occupation, places, etc.), so we do not take into account the number of markers, definitive articles, conjunctions, etc.

C. Results and Discussion

Some qualitative results can be seen in Figure 1. There are several aspects to remark: First, it is not necessary to remove the framing borders in black, and even the white section at the bottom left corner does not affect to the final result. Second, the dark vertical stripe in the middle of the right side page has almost disappeared. Third, even when the verso components are very dark (see Fig.1c), the show-through phenomenon is significantly reduced (see Fig.1d), improving the readability of the document.

Quantitative readability results are shown in table I. From the 22 pages, the scholar reported at step 1 that there were 33 unreadable *important words*. At step 2 we provided the enhanced images, and the expert retrieved a total of 25 *new words* (hence, 8 remained unreadable). The scholar stated that in 3 of these 8 remaining unreadable words, the problem was not the quality of the restoration but the understadability of the handwriting style. Hence, the total amount of recovered words can be considered 28 (instead of 25). According to these results, the recovery rate of unreadable words is 85%.

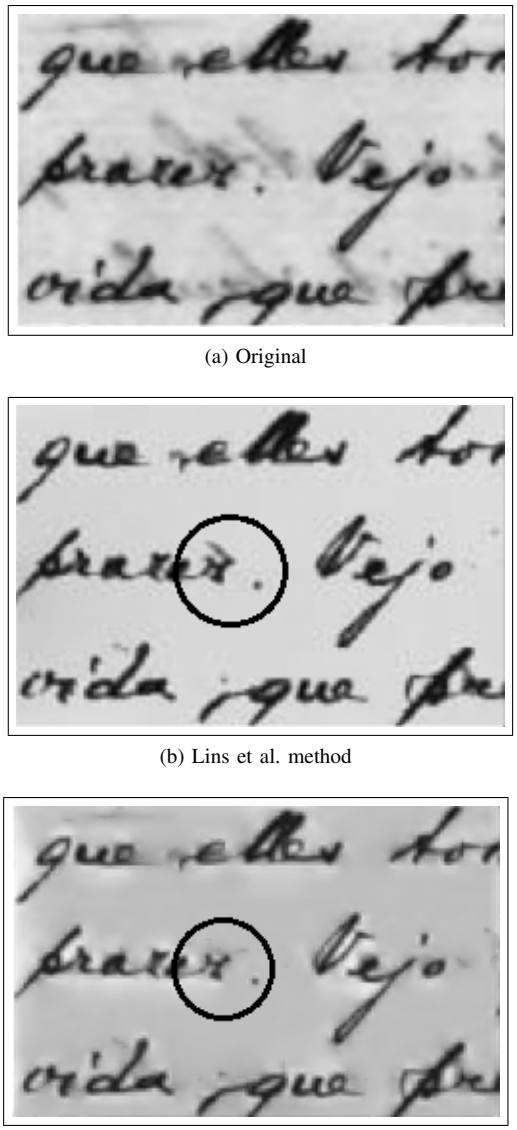
TABLE I: Readability experiments performed on the 22 pages corresponding to the old marriage records in microfilm format.

Unreadable words	Recovered words	Recovery rate
33	28	28 / 33 = 85%

Independently of the 33 initially unreadable words, the scholar reported 11 *modified words* that were previously considered as readable and correct. These modified words are very important for scholars, since it is of key importance to have trusthful information.

The scholar also stated that, with the improved images, the time spent for reading every page could be significantly reduced, saving an important human effort.

As we have comented before, it is difficult to quantitatively evaluate the quality of show-through removal methods. For this reason, next we will show some qualitative results using document images that have been used for evaluating other existing methods. In Figure 2 and Figure 3, we show some examples from Lins et al. [10] and Nishida and Suzuki [9]. Since the input of our method is a grey-scale image, we have converted the original color image into grey-scale before applying our algorithm.



(a) Original

(b) Lins et al. method

(c) Our method

Fig. 2: Show-through removal example using the images extracted from Lins et al. in [10]. Notice that we can correctly recover the letter 'r' that is shown inside the circle in black color.

Figure 2 shows the restored images. Notice that, contrary to Lins' method [10], we can perfectly recover the letter 'r'. However, in some cases our enhanced output image contains some visible white artifacts. It is known that hard thresholding of wavelet coefficients produce visible artifacts [14]. This effect can be seen on Figure 2c where some white artifacts appear around letters.

Figure 3 shows another example of a restored image. Notice that, in contrast to the method presented in [9], we can correctly remove the long verso horizontal wide stripes, although we loose some of the foreground dots at the bottom left corner.

(a) Original (b) Nishida and Suzuki's method (c) Our method

Fig. 3: Show-through removal example extracted from Nishida and Suzuki's method [9].

IV. Conclusions

In this paper we have proposed a novel method for enhancing and canceling the show-through phenomenon in grey-scale documents. The method only requires one side of the page, does not need to remove the framing borders, and only one parameter has to be tuned. The experiments have shown that the resulting images improve the readability of the documents, allowing scholars to recover unreadable words, to solve ambiguities and, of course, to reduce the amount of time spend for reading every document.

It must be said that the results obtained with the proposed readability metrics could be argued as being subjective. Although we have validated our method using the opinion of an expert on medieval documents, we are aware that there is a lack of trustworthy ground-truth for the microfilms. Therefore, no one can ensure the real content of these degraded documents (for example, the scholar can consider he correctly readed *Smith* when the true word is *Smit*). For this reason, we plan to test our method using some public databases.

Further work will be focused on automatically setting the threshold parameter, proposing a soft thresholding method to decrease the visible artifacts and using filters to remove the blur. We will also investigate the extension of the method in order to deal with color documents.

Acknowledgment

We would like to thank the Cathedral of Barcelona, the Center for Demographic Studies (UAB), and Miquel Amengual (expert on medieval documents) for participating in the readability experiments. This work has been partially supported by the European Research Council Advanced Grant (ERC-2010-AdG 20100407: 269796-5CofM) and the Spanish projects TIN2009-14633-C03-03, TIN2012-37475-C02-02 and TIN2010-21771-C02-01.

References

[1] J. da Silva, R. Lins, F. Martins, and R. Wachenchauzer, "A new and efficient algorithm to binarize document images removing back-to-front interference," *Journal of Universal computer science*, vol. 14, no. 2, pp. 299–313, 2008.

[2] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A combined approach for the binarization of handwritten document images," *Pattern Recognition Letters*, vol. online first, 2012.

[3] ——, "Performance evaluation methodology for historical document image binarization," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 595 –609, feb. 2013.

[4] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The esposalles database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, no. 6, pp. 1658–1669, 2013.

[5] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 736–754, 2001.

[6] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *International Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 17–25, 2007.

[7] F. Merrikh-Bayat, M. Babaie-Zadeh, and C. Jutten, "Linear-quadratic blind source separating structure for removing show-through in scanned documents," *International journal on document analysis and recognition*, vol. 14, no. 4, pp. 319–333, 2011.

[8] Q. Wang, T. Xia, L. Li, and C. Tan, "Document image enhancement using directional wavelet," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2003, pp. II–534.

[9] H. Nishida and T. Suzuki, "A multiscale approach to restoring scanned color document images with show-through effects," in *International Conference on Document Analysis and Recognition*, 2003, pp. 584–588.

[10] R. Lins, J. Silva, S. Banergee, A. Kuchibhotla, M. Thielo *et al.*, "Enhancing the filtering-out of the back-to-front interference in color documents with a neural classifier," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 2415–2419.

[11] G. P. Silva, R. Lins, and J. Silva, "Histdoc-a toolbox for processing images of historical documents," in *Proceedings of the 7th international conference on Image Analysis and Recognition-Volume Part II*. Springer-Verlag, 2010, pp. 409–419.

[12] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America. A, Optics and image science*, vol. 7, no. 10, pp. 2032–2040, 1990.

[13] R. Lins, J. da Silva, and F. Martins, "Detailing a quantitative method for assessing algorithms to remove back-to-front interference in documents," *Journal of Universal Computer Science*, vol. 14, no. 2, pp. 266–283, 2008.

[14] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994. [Online]. Available: http://biomet.oxfordjournals.org/content/81/3/425.abstract